THE ULTRA HIGH DENSITY STORAGE OF NON-BIOLOGICAL INFORMATION IN A MEMORY COMPOSED OF DNA MOLECULES

Junghuei Chen*, and Yuzhen Wang
Chemistry & Biochemistry, University of Delaware
Newark, Delaware, 19716

Russell Deaton
Computer Science & Computer Engineering, University of Arkansas
Fayetteville, Arkansas, 72701

ABSTRACT

We have designed and performed a proof-of-principle experiment demonstrates that huge amount of non-biological, or abiotic, information can be stored in a memory composed of DNA molecules. The preliminary experiment emphasizes on achieving a practical design motivates several fundamental questions, such as the amount of information that can be stored in a DNA memory before errors are introduced, and practical and cost-effective ways of mapping abiotic data onto DNA sequences.

1. INTRODUCTION

The DNA memory implementation consists of two major stages that progressively add capability. The first stage is called a Read/Write (R/W) DNA memory, in which practical storage and accurate retrieval of information are demonstrated. Here, with a library of 800 words we created and store a database of customer information with DNA. We also performed a search and retrieve operation with the DNA database. Finally, we showed that with a divide-and conquer method we can read out the retrieved information with a DNA microarray containing the 800 words library.

1.1 Store Abiotic Information in DNA

As illustrated in Figure 1, we first constructed a word library with random 40-mer DNA sequences by cloning the PCR product of this library into plasmids. After transforming it into *E. coli* host cell, we generate a collection of E. coli strains each containing an unique 40-bp sequence that can be amplified and isolated by cutting it out with two built in restriction enzyme sites flanking the 40-mer sequences. Next, we assigned one word for each unique 40-mer sequence (less than 50,000 words are needed for practical communication in English). Then, we connected the words to a sentence or entry of short information independently with DNA ligase. The sum of all the sentences or entries is our database; small samples shown at the bottom of Figure 1.

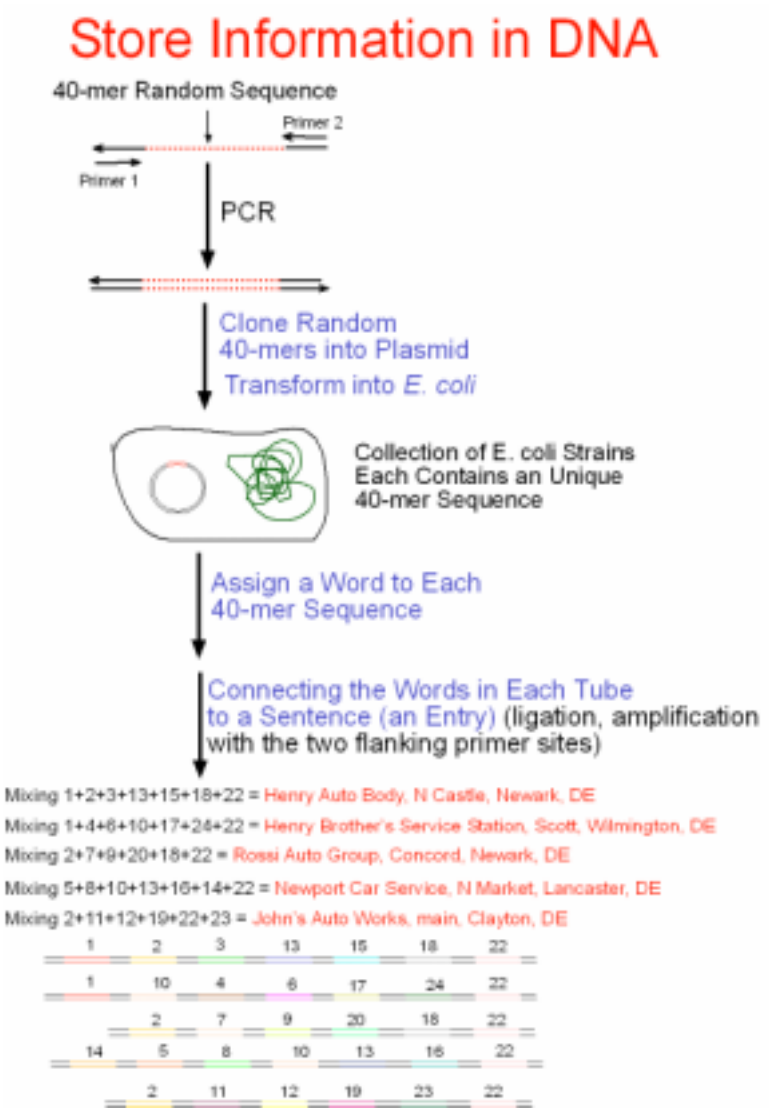

Figure 1

1.2 Retrieve and Read-out of the Stored Information

To retrieve and read the stored information, we first pass the DNA database through an ssDNA column with the sequence of the word of interest. The retrieved DNA (sentences or entries by Watson-Crick hybridization)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Ultra High Density Storage Of Non-Biological Information In A Memory Composed Of Dna Molecules	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chemistry & Biochemistry, University of Delaware Newark, Delaware, 19716; Computer Science & Computer Engineering, University of Arkansas Fayetteville, Arkansas, 72701**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida. , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**2**	

can be used as the probe on a DNA microarray, in which each spots represents one word. The contents of the retrieved sentences can be read out from the DNA microarray reference. The retrieved DNA can be further separated according to another word by passing through another column until a single sentence or entry can be read out from the DNA microarray (see Figure 2).

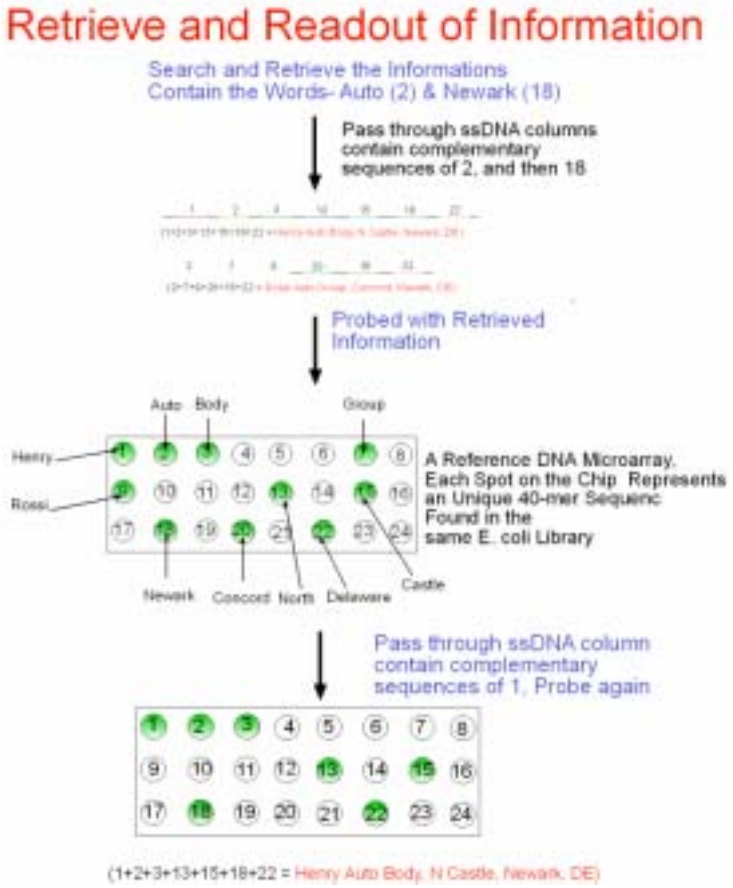

Figure 2

2. DISCUSSION

2.1 Why DNA?

Advances in information technology have produced vast amounts of data. A study by the University of California at Berkeley in 2000 (Lyman et al., 2000) estimates that 1 to 2 exabytes (10^{18} bytes) of information are produced worldwide, each year. In addition, the amount of information that is published in periodicals and books doubles every few years (Wurman, 1989), and the Internet now has search engines with 3 billion URLs (Google, 2003). As the amount of data continues to increase, the ability to store it and search for relevant information is diminished. DNA has many attractive properties as a storage medium outside of the cell. A gram of DNA (can be easily resolved in a volume of one drop of water) potentially can store ~ 10^{21} bits of information.

To take an example, there are approximately 500,000 words in the Oxford English Dictionary (OED). There are 4^{20} (~10^{12}) DNA molecules of length 20 base pairs (bp). Thus, the content of the OED is small fraction of the total number of 20-mers, and could be easily stored in a DNA memory with an appropriate word to sequence mapping. Moreover, DNA is a molecule on the nanoscale that can be manipulated in the test tube to read and write data with well-characterized molecular biology protocols as we have demonstrated here. DNA can be attached to other nanomaterials, such as carbon nanotubes (Williams et al., 2001) and gold nanoparticles (Mirkin et al., 1996), and has been used to assemble nanostructures from those materials (Mirkin et al., 1996). Small amounts of DNA can store vast amounts of information. The operations on DNA molecules occur in parallel, which produces substantial speed-ups for very large data sets. Moreover, DNA computing is capable of universal computation (Paun, 1996), and computations could be done in vitro on the information stored in the memory.

REFERENCES

Lyman, P., Varian, H. R., Dunn, J., Strygin, A., and Swearingen, k., 2000: How much information? http://www.sims.berkeley.edu/research/projects/how-much-info/.

Wurman, R., 1989: Information Anxiety. New York: Doubleday.

Google, 2003: http://www.google.com.

Williams, K. A., Veenhuizen, P. T. M., d. Torre, B. G., Eritja, R., and Dekker, C., 2001: Carbon nanotubes with DNA recognition. *Nature*, **420,** 761-766.

Mirkin, C., Letsinger, R. L., Mucic, R. C., and Storho, J. J., 1996: A DNA-based method for rationally assembling nano-particles into macroscopic materials. *Nature*, **382,** 607-609.

Paun, G., 1996: Universal DNA computing models based on the splicing operation, in Landweber and Baum pp. 59, 76. DIMACS Workshop, Princeton, NJ.

CONCLUSION

With a huge capacity, and massively parallel search abilities, a DNA memory for abiotic data is a potentially revolutionary way of storing and processing vast amounts of information. In the future, large scale data storage and a limited intelligence will be added to the R/W DNA memory in the form of context-based search. Thus, it should be termed the Intelligent DNA memory. For example, the entire database of customer information from Acxiom Corporation (a database service company) can be stored, searched, and processed in DNA form. The Acxiom database is approximately 6 Petabytes (~10^{15} bytes that is stored in a warehouse of tape and disk farm half the size of a football field) can be stored in a volume less than a drop of water. For the Acxiom application, it is not enough to store and recall information, but because of imprecision in the customer data, such as misspellings, abbreviations, and synonymous terms, searching and retrieving data based upon both content and context is advantageous.